# Data Wrangling Project: Sharks Attack Dataset

## Instructions

Welcome to the final project of this data wrangling module! In this project, you will get a chance to work through the entire data wrangling workflow while preparing the shark\_attacks.csv file for analysis. This dataset contains very dirty data and will require a lot of work! This project is broken down into key steps of the data wrangling process to help guide you along the process. When you are finished, save the wrangled dataset as a final\_project.csv file. Submit the final project as a zip folder named final\_project.zip. Make sure the zipped folder has both your wrangled dataset and this word document within it. Best of luck!

## Step 1: Decide which tool to use

This dataset contains around 1100 rows. Discuss which tool (BigQuery/Python/Google Sheets) is best suited for the data cleaning task for this dataset. Mention the relevant advantages and disadvantages of each tool. Finally, state which tool you think is best suited for the task and why. (6 marks)

|  |
| --- |
| <need to amend>  **Google sheets:**   * Best for: Quick data exploration, simple calculations, basic data visualization, and small to medium-sized datasets. * Advantages: User-friendly interface, no programming required, and suitable for non-technical users. * Limitations: Limited scalability, not ideal for complex data transformations or large datasets.   **Python:**   * Best for: Advanced data analysis, data cleaning, manipulation, visualization, machine learning, and working with large datasets. * Advantages: Versatile, extensive libraries (e.g., Pandas, NumPy, Matplotlib, Seaborn), automation, and reproducibility. * Limitations: Requires programming skills, steeper learning curve for beginners.   **SQL (Structured Query Language):**   * Best for: Data retrieval, filtering, aggregation, and working with structured databases. * Advantages: Efficient querying, suitable for large datasets, essential for database management, and often used in conjunction with other tools like Python. * Limitations: Focused on database operations, not ideal for complex data transformations or advanced analytics.   Python, and Excel |

## Step 2: Data Inspection

Inspect the dataset. In the box below, discuss the following:

* Are there any irrelevant columns? Which ones?
* Are there any duplicates?
* Which columns have missing data?
* For each column with missing data, describe what you think the best way to handle that missing data is, and why?
* Are there any errors? Describe any you find.
* Is there anything else that requires data cleaning attention?

(12 marks)

|  |
| --- |
|  |

## 

## Step 3: Data Cleaning

Following on from Step 2, clean the dataset. Document all the changes you make in the box below. Before data cleaning, make sure to check every column thoroughly (audit the data). List all the actions to take so that you don’t overlook anything. (12 marks)

|  |
| --- |
|  |

## 

## Step 4: Data Cleaning Validation

Go through the data cleaning checklist and make sure there is no dirty data remaining! List below all the data validation steps you take. (3 marks)

|  |
| --- |
|  |

## Step 5: Data Enrichment

With the dataset cleaned it’s time to enrich the data:

* Make an address column, by combining the Location, Area and Country columns together (this might affect your missing value strategy!).
* Add a new column, call it “Shark”. Extract information from the Species column. If the species text mentions the word “white”, make the “Shark” column value “Great White”. If the text mentions “bull”, make the “Shark” column value “Bull”. Otherwise, if neither of the words found, make the value “Other”. (Hint: make sure the species column is all lowercase).

## Step 6: Publish the dataset

Export the data as csv file. Call it final\_project.csv. Submit the file in a zip folder called final\_project.zip. Make sure the zip folder contains both your wrangled dataset and this word document with your answers!

## 